**Case Study: Brest Cancer prediction**

**Introduction**

According to cancer.org, breast cancer (BC) is common cancer among women worldwide. There is a 1 in 8 chance that a woman will develop breast cancer, representing most new cancer cases and cancer-related deaths, making it a significant public health problem in today’s society.

The early diagnosis of BC can significantly improve the prognosis and chance of survival, as it can promote timely clinical treatment. The accurate classification of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classifying patients into benign or malignant groups is the subjected to extensive research. Data mining and machine learning (ML) is widely recognized as the methodology of choice in breast cancer pattern classification and forecast modeling because of its unique advantages in critical features detection from complex BC datasets.

Classification and data mining methods are effective ways to classify data, especially in the field of clinical analysis, these methods are widely used in diagnosis and analysis to make decisions.

My solution to this problem involves building a model that accurately classifies tumors as Benign or Malignant based on the tumor shape and geometry.

**About Data:**

This is Dataset classifies the Benign and Malignant cells using the description about the cells in the form of columnar attributes. This data was donated by researchers of the University of Wisconsin and includes the measurements from digitized images of fine-needle aspirate of a breast mass.

The data sets is provided with 569 examples of cancer biopsies, each with 32 features. One feature is an identification number, another is the cancer diagnosis and 30 are numeric-valued laboratory measurements.

The diagnosis is coded as

* "M" to indicate malignant
* "B" to indicate benign.

The other 30 numeric measurements comprise the mean, standard error and worst (i.e. largest) value for 10 different characteristics of the digitized cell nuclei

Ten real-valued features are computed for each cell nucleus

* radius
* texture
* perimeter
* area
* smoothness
* compactness
* concavity
* concave points
* symmetry
* fractal dimension

For each image - The mean, standard error and "worst" of these features were computed and this results in 30 features.

For example; the real-vales feature radius is computed as

* radius\_mean
* radius\_se
* radius\_worst

There are no missing values identified in dataset.

Class distribution: 357 benign, 212 malignant

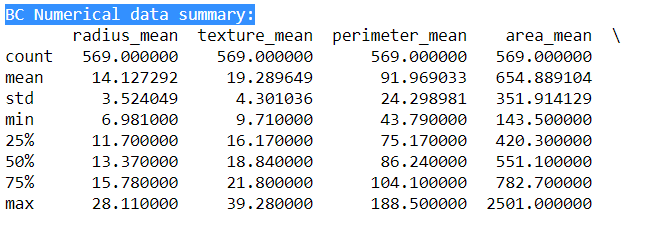
**Graph Analysis**

**Data cleansing steps:**

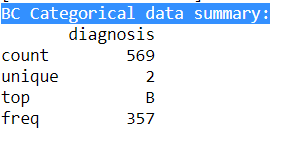
* Dropped ID column – this is not required for analysis.
* Checked not null count. Not null count 569 is matching with RangeIndex 0 to 568. This proves that data does not contain nulls.
* The dimension of the table is: (569, 31)

**Summary of numerical data and categorical data**

**BC Numerical data summary:**

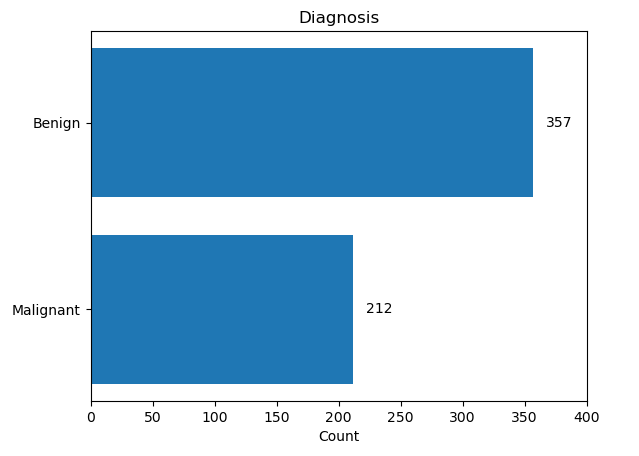


**BC Categorical data summary:**



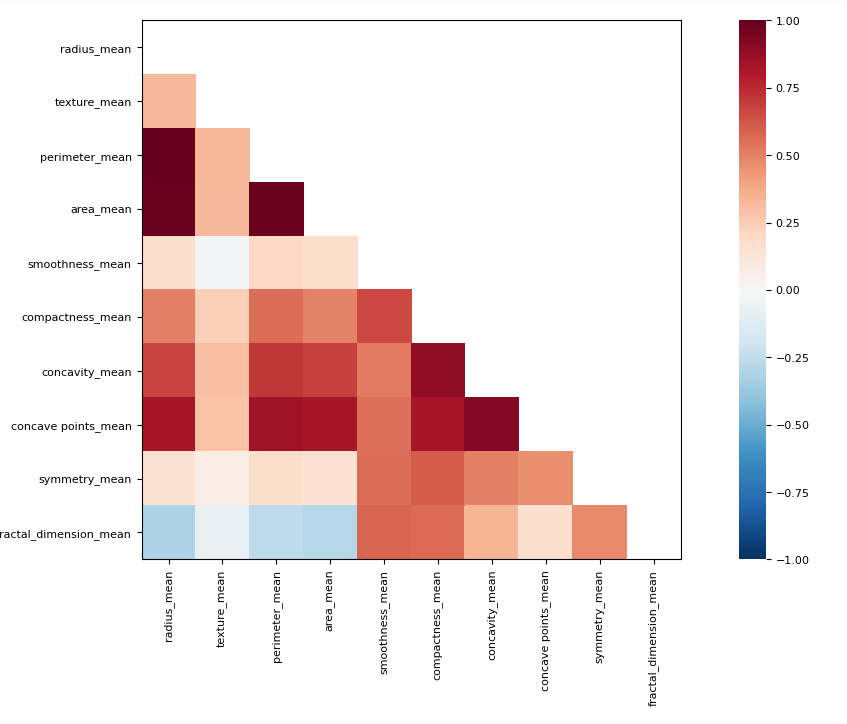
**Bar Chart:**

Below chart represents the break up count of malignant and benign



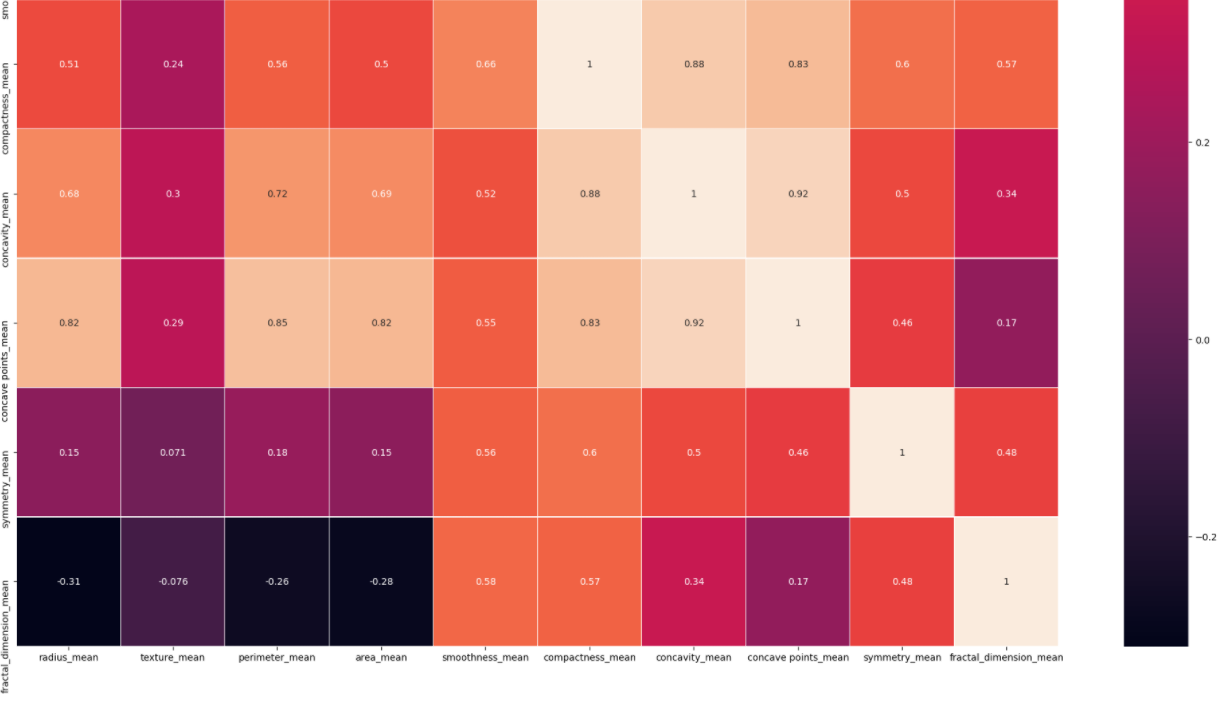
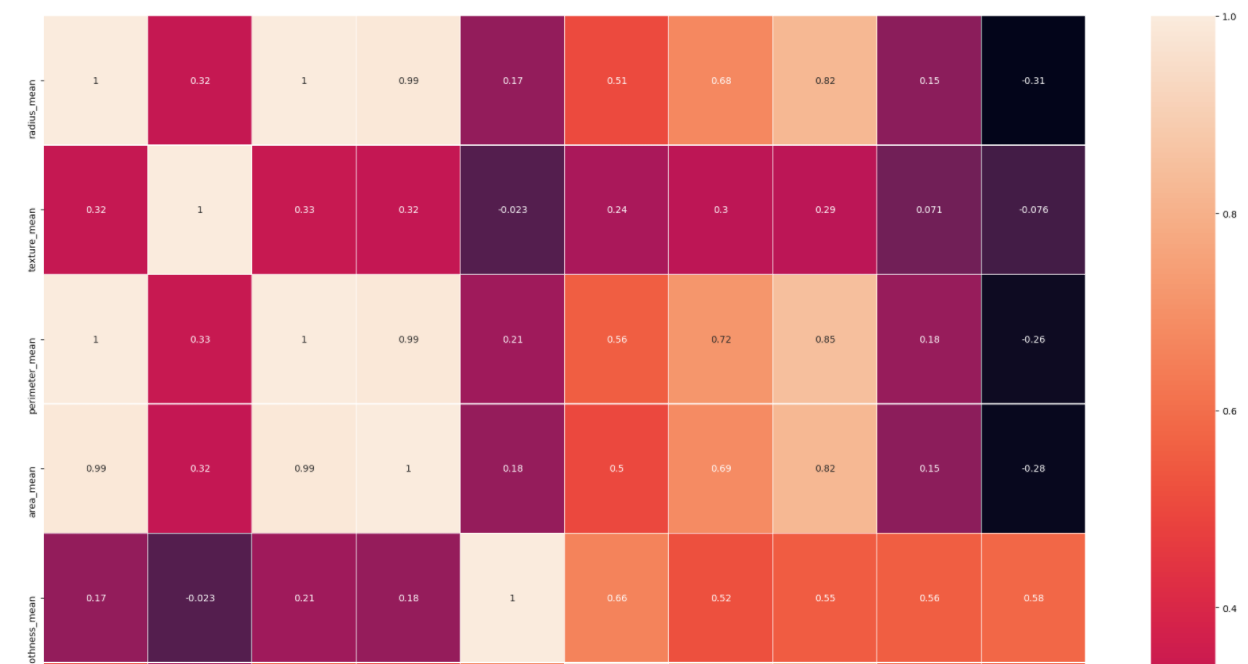
The dataset contains a class imbalance between the malignant and benign cases. Out of the 569 cases in the dataset, 62%(count: 357) cases are benign cases of cancer whereas 37% (count 212)cases are malignant.

**Correlation plot** is demonstrated using Pearson Ranking, Heat Map and pair plot:

An important step in the graph data analysis step is to identify if at all there is any correlation between any of the 32 variables. First, I have considered 8 mean variables; by using Pearson’s correlation, I have identified that the positive correlation between two variables is demonstrated through the darkness of the blue color, i.e., the darker the blue-colored box, the stronger is the positive correlation between respective variables. Similarly, a negative correlation between two variables is demonstrated through the darkness of the orange color, i.e., the darker the orange-colored box, the stronger is the negative correlation between respective variables.

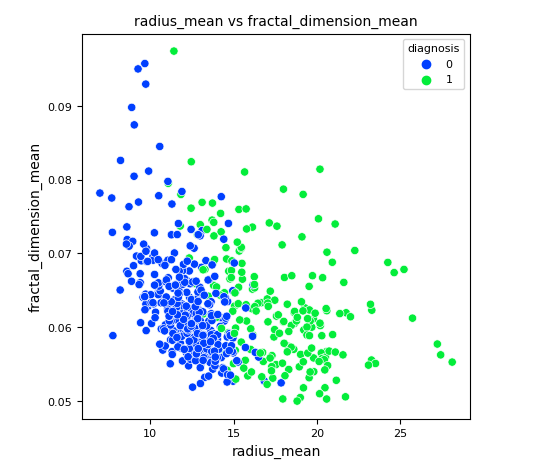
**Heatmap:**

The positive correlation between two variables is demonstrated through the darker black shade color, i.e., the darker the black shade, the more substantial the positive correlation between respective variables. For example, radius\_mean vs. fractal\_dimension\_mean have a positive correlation. Similarly, the negative correlation between two variables is demonstrated through the lighter orange, i.e., lighter the colored box. More substantial is the negative correlation between respective variables, such as radius\_mean vs. perimeter\_mean.

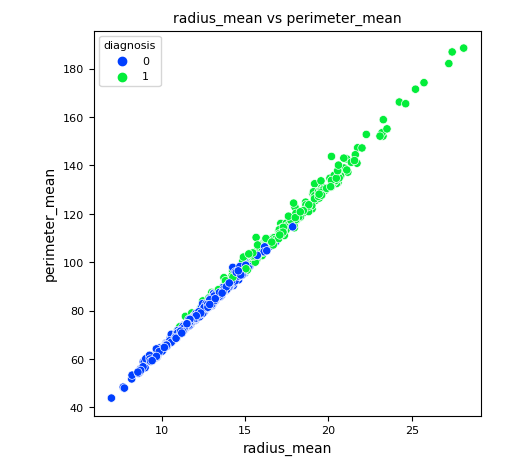


**Scatter plot**: From the heatmap and Pearson Ranking it’s been observed that the highly correlated pairs are:

Positive correlation: radius\_mean vs fractal\_dimension\_mean



Negative correlation: radius\_mean vs perimeter\_mean

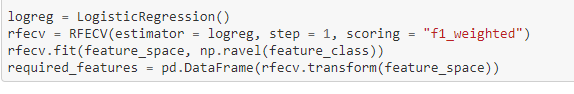


**Dimensionality & Feature Reduction and Feature Engineering**

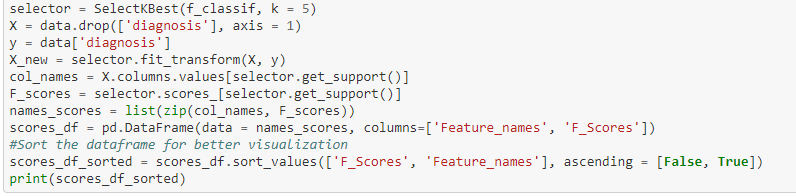
Dimensionality Reduction Dimensionality Reduction is a technique that reduces the number of independent variables to a collection of theory variables by eliminating certain values that are less important in predicting the results. There may be many independent variables, but we have two independent variables at last by applying a suitable dimensionality reduction technique. There are two processes, collection of Feature and Extraction of feature. It is basically used to get two-dimensional data so that a better visualization of machine learning models can be done by plotting the prediction regions.

Below are the steps performed to select top five features for the model.

* Fetch the required features using RFECV - Recursive Feature Elimination and Cross-Validation Selection
* With the help of RFECV we can eliminate the irrelevant features based on scoring.
* I have used “f1\_weighted” for scoring - It results in an F-score that is not between precision and recall



F\_score to pick the top five features:



**Result:**

**Feature\_names** **F\_Scores**

4 concave points\_worst 964.385393

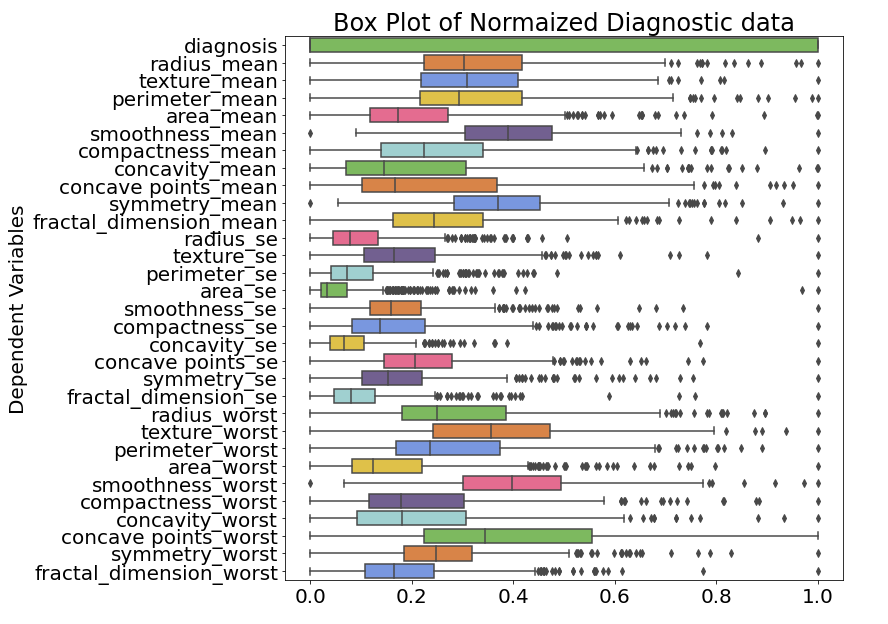
3 perimeter\_worst 897.944219

1 concave points\_mean 861.676020

2 radius\_worst 860.781707

0 perimeter\_mean 697.235272

**Box plot :**

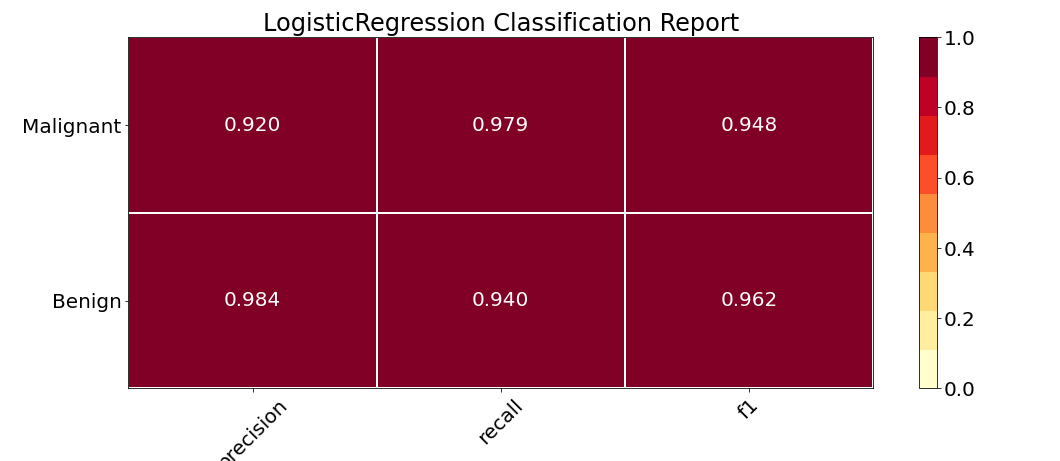


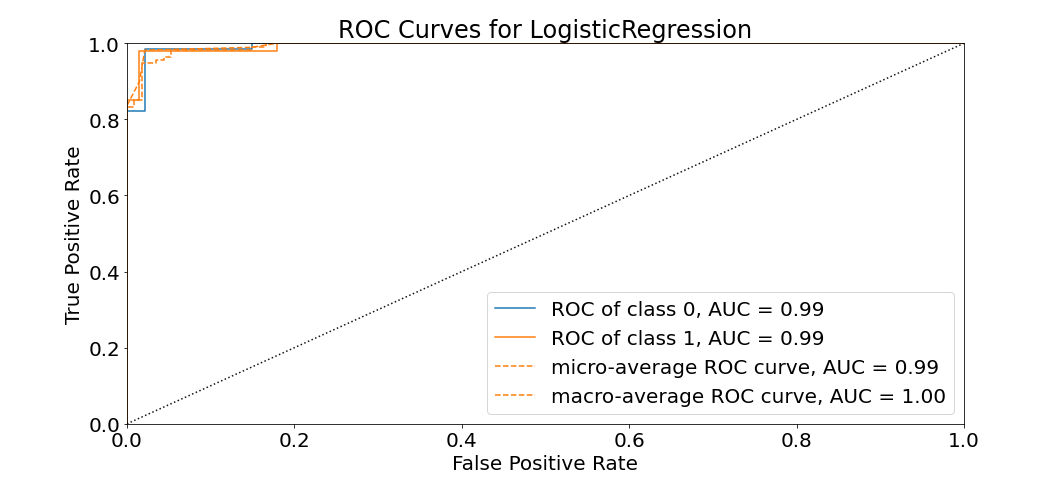
SelectKBest proves that the top 5 features are not contributing to eliminated features using RFECV and box plots. Hence it is safe to eliminate the features area\_mean,concave points\_se,area\_worst

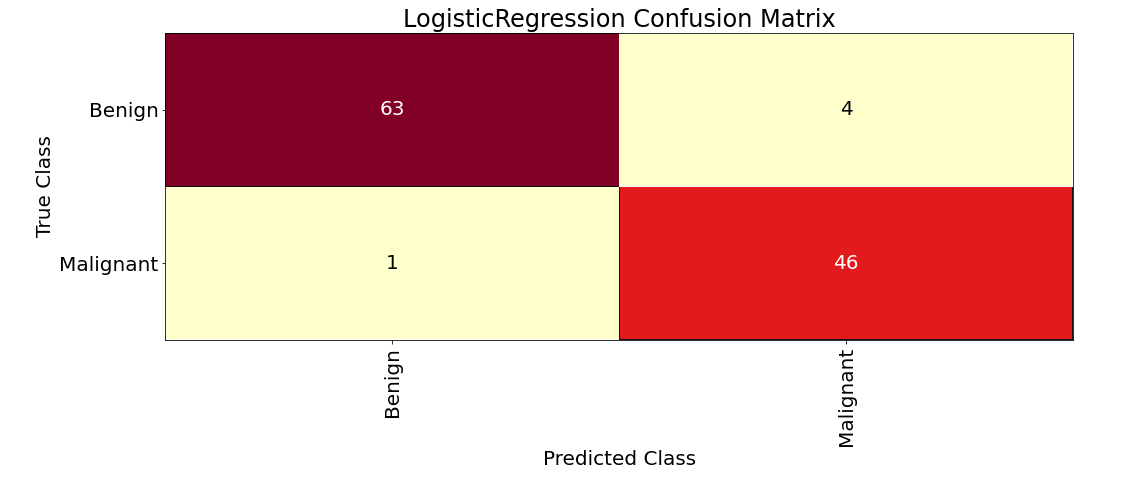
**Model Selection & Evaluation**

I would be using Logistic Regression, Keras Classifier, Support Vector Machine(SVM)

**Logistic Regression:**

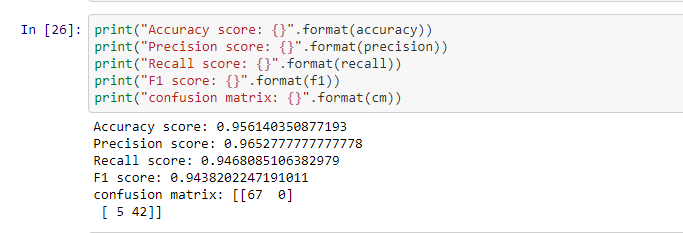


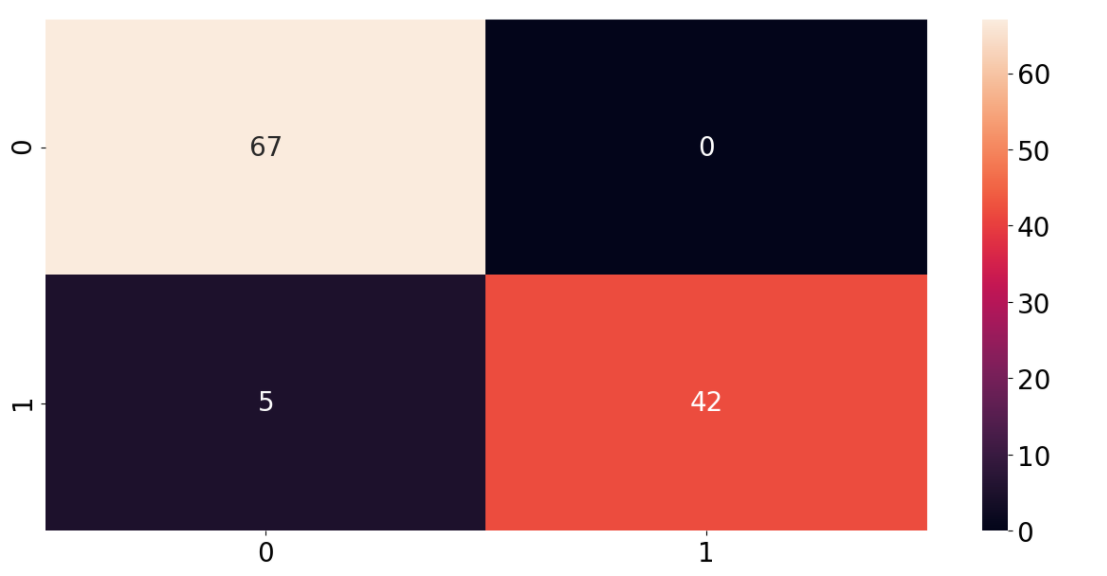




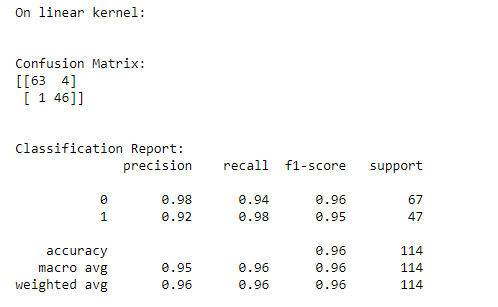
Formula (TP+TN)/(TP+TN+FP+FN)= (63+46)/(63+46+1+4)= 109/114= 0.95 i.e 95 % of accuracy.

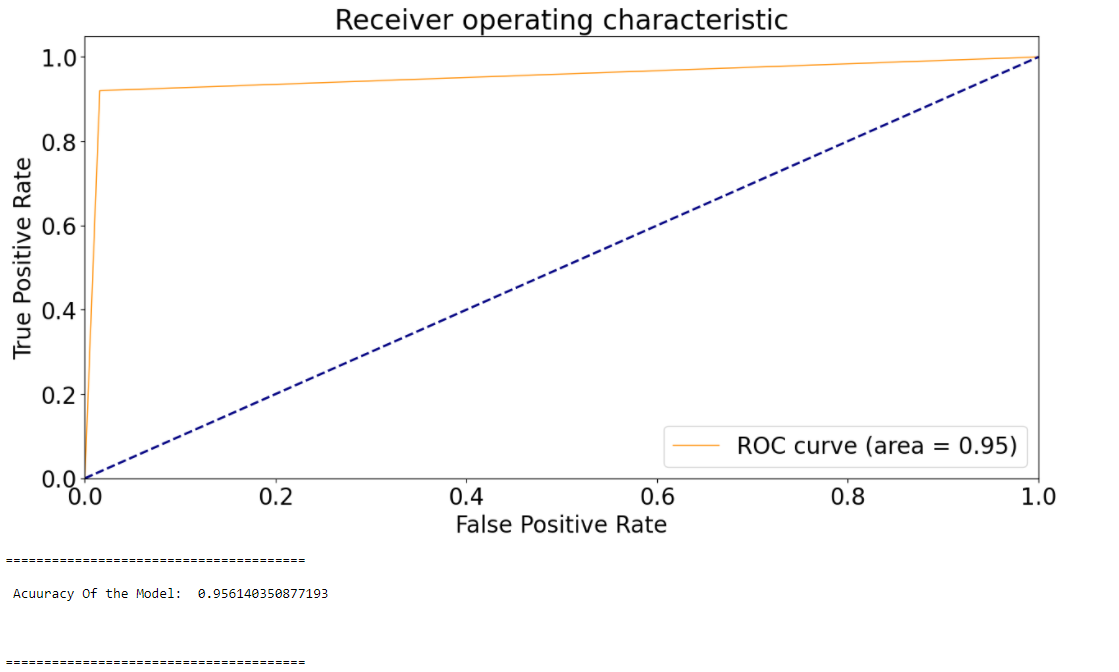
# Keras Classifier





# Support Vector Machine(SVM)





**Key Observations:**

**Logistic Regression**

1. Accuracy score: 0.96
2. Precision score: 0.98
3. Recall score: 0.94
4. F1 score: 0.96
5. ROC curve area 0.99

**Keras Classifier**

1. Accuracy score: 0.96
2. Precision score: 0.97
3. Recall score: 0.95
4. F1 score: 0.94

**Support Vector Machine**

1. Accuracy score: 0.95
2. Precision score: 0.92
3. Recall score: 0.98
4. F1 score: 0.95
5. Roc Curve area 0.95

The Roc curve area is good at logistic regression compared to SVM model. Considering the fact that the Accuracy score is almost same for all the models. Precision is slightly varied but it is significant among the models. Similar to Precision F1 Score has significant difference.Recal Score is high for SVM when compared to Logistic regression.

**Conclusion:**

In this project, I have created three classification machine learning models that can predict if a person has breast cancer based on 10 different characteristics of the digitized cell nuclei. In comparison between logistic regression, Keras and SVM, the logistic regression model was more accurate in predicting breast cancer’s class. Therefore, it seems that for classification of breast cancer’s class, logistic regression method is appropriate to be used. The Logistic regression, correctly classifies patients with and without breast cancer 96% of the times. Its AUC of 99% indicates a great ability to distinguish between a benign lump and a malignant tumor.

**References:**

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Anand, A. (2020, December 20). *Categorical feature selection using chi- squared test*. Medium. <https://medium.com/analytics-vidhya/categorical-feature-selection-using-chi-squared-test-e4c0d0af6b7e>.

Raj, J. T. (2019, March 14). *Dimensionality reduction for machine learning*. Medium. <https://towardsdatascience.com/dimensionality-reduction-for-machine-learning-80a46c2ebb7e>.